Student perceptions of four university gateway tests

JOHN GARDINER

University of Sydney Centre for English Teaching

STEPHEN HOWLETT

University of Sydney

ABSTRACT

By listening to test-takers 'stories' behind standardised English tests, educators can better prepare and advise students taking one of four university gateway tests (IELTS, PTE-A, CAE and TOEFL iBT). This study presents a comparative analysis using qualitative methodology of the perceptions of 25 English as a second language (ESL) students in a 25-week direct entry course to university. One week prior to sitting each of the four tests, participants were offered two instruction classes for familiarisation purposes. Through several sources of data-gathering, comprising questionnaires and interviews before and after taking each test, and two focus group discussions during their first university semester, themes were coded and analysed to better understand the test-taker experiences and test challenges. The results indicated that each test has perceived differences that need to be specifically addressed. The research also found that test item and test format familiarity were crucial to reduce anxiety and improve test score validity. The results provide insight into the perception of students who have taken all four university 'gateway' tests within a two-month period. The implications of student voice in this context are relevant to test preparation course developers, educators and other test stakeholders.

Address for correspondence: John Gardiner, University of Sydney Centre for English Teaching, Room 525, Level 5, Wentworth Building G01, Darlington NSW 2008; Email: john.gardiner@sydney.edu.au

University of Sydney Papers in TESOL, 11, 67-96.

 ISSN 1834-3198 (Print); ISSN: 1834-4712 (Online)

INTRODUCTION

The language readiness of prospective international students for tertiary study is currently determined through high-stakes standardised English proficiency tests. O'Loughlin (2013) points out that English proficiency decisions at Australian universities are mostly based on a single standardised test score. Even though this dependence on a sole determinant has been questioned (Dunworth, 2010; O'Loughlin, 2011), entry requirements for international students remain unchanged. The test-types accepted at many Australian universities now include four standardised tests, The International English Language Testing System (IELTS), the Pearson Test of English - Academic (PTE-A), Cambridge English Advanced Exam (CAE), and Test of English as a Foreign Language internet-Based Test (TOEFL iBT). Most studies investigating test-taker experiences have attempted to explore different phenomena, rather than trying to replicate the phenomena from previous studies. Areas of test-taker research include test preparation (Moore, 1994; Green, 2007; Sadeghi, 2014; Liu, 2014), 'washback', or test impact on teaching and learning (Alderson & Wall, 1993; Shohamy, 2001; Cheng, Sun& Ma, 2015), test-taking strategies (Cohen, 2013; Matoush & Fu, 2012), attitudinal or psychological factors (Stricker & Attali, 2010; Ata, 2015; Stankov, Lee, Luo & Hogan, 2012), single skill analysis (writing, He & Shi, 2008; lexical factors, Webb & Paribakht, 2015) and, test mode factors (Bernstein, Van Moere & Cheng, 2010; Barkaoui, 2014).

While research into teacher test preparation activities gained attention in the 1980s and 1990s (e.g. Moore, 1994; Popham, 1991), few studies (e.g. Sadeghi, 2014; Yu, 2012) have been conducted since then and they tend to provide mixed results. Although some studies (Green, 2007) conclude that preparation focusing only on test-taking strategies return no substantial benefit, others (Elder & O'Loughlin, 2003; Liu, 2014) report a slight benefit or even a considerable benefit (McNeil, 2011) in terms of score improvement. Alderson and Wall (1993) published the seminal text often credited as laying the foundation for 'washback' research. Both Green (2013) and Cheng et al., (2015) have reviewed the literature in this area since 1993, noting that studies on learners have only started to appear. Green (2013) synthesised these studies, concluding that content rather than teaching methods tend

to change to reflect test tasks. In relation to test-taking strategies, it is important to differentiate 'test-management' skills from 'testwiseness' skills. Cohen (2013) refers to "test-management" (p. 895) as strategies consciously selected for responding meaningfully to test items and tasks. In other words, 'testwiseness' only requires test-taking techniques, whereas implementing 'test-management' requires the development of language and cognitive skills.

Attitudinal or psychological factors have been investigated in relation to test-taking. Test familiarity and test anxiety have been linked in the literature. Winke and Lim (2014) concluded that unfamiliarity with a test can cause heightened test-taker anxiety and poor "test-management" skills. Suryaningsih (2014) reached a similar conclusion in a study of TOEFL and IELTS test-taker perceptions by stating that test format and rubric familiarity affected their performance. A Singapore study on confidence (Stankov et al., 2012) provided empirical evidence that confidence is a strong predictor of success, perhaps more so than self-efficacy and anxiety. In relation to teaching, they indicate that confidence in choosing a wrong answer has become a valuable information source for remedial science classroom activities, implying a possible transfer to other subjects. In a study of attitudes to the TOEFL iBT test, Stricker and Attali (2010) observed that "relevant data about current test-taker attitudes are sparse,"(p.14), highlighting the research gap in this area.

Studies have also investigated single language skills such as writing. He and Shi (2008) reported on Chinese student perceptions of two standardised English proficiency writing tests at a Canadian university. They observed that cultural differences regarding preparation content and teaching expectations caused considerable disappointment among students expecting memorisation exercises "like filling in the blanks" on general structures or generic sentence samples (p.137). This has pedagogical implications for educators with students from culturally diverse educational backgrounds. In a lexical profile of reading and listening in standardised tests, Webb and Paribakht (2015) suggest that developing vocabulary knowledge may improve standardised English test performance. Test mode studies on computer-based tests such as Pearson's PTE-A and TOEFL iBT often relate to speaking and writing sections. In one correlation study, (Bernstein et al., 2010) examined

whether PTE-A automated speaking tests were as valid as human-rated tests, indicating that automated test scores strongly correlated with oral proficiency scores. However, research into task authenticity and idea coherence is limited. Stricker and Attali (2010) noted in their study of student perceptions of the TOEFL iBT across four countries that the speaking section was less favourable in all countries, but were unable to determine the reasons. Another area of test-mode research concerns the effect of keyboarding skills on validity. In a study of 97 TOEFL iBT test-takers, Barkaoui (2014) provides evidence that poor keyboarding skills had a significant but weak effect on task scores.

The investigation of different phenomena and methods is an issue in the test-taker literature. However, some researchers (Song & Cheng, 2006; Cheng & DeLuca, 2011) have recognised this problem and attempted to link these disparate studies to test reliability or validation. Building on attempts to develop a more systematic and rigorous classification of data, overlapping themes and sub-themes in four studies on test-taker perspectives were analysed. This study contributes additional understanding to that reported by Cheng and DeLuca, 2011; DeLuca, Cheng, Fox, Doe and Li, 2013; Puspawati, 2012; Suryaningsih, 2014; and others, of perceived test-taker experiences of four university gatekeeper English proficiency tests. To our knowledge, it is the first study of this kind in which students who have taken all four gateway tests within a short time period have voiced their experiences. It further offers new insight into types of test items and skills likely to be problematic for international students in these tests. Teaching implications of areas identified as requiring extra attention and support from teachers will also be discussed.

Research Questions

This paper presents a qualitative report on international students' perceptions of the English proficiency test-taking experience by addressing the following:

1. How do students perceive their experiences of taking four university 'gateway' English proficiency tests?
2. What test-item types and skill types do test-takers perceive to be more challenging across these four high-stakes tests, thus requiring more preparation?

METHOD

The research design and procedures used in this paper were informed by TEFL industry feedback from several colloquiums and TESOL workshops. The intention was to consider all stakeholders in this study, by including university human ethics approval and test-provider considerations. As it is not the intention of the paper to rank the four tests, each test is described in terms of relevant themes. The study draws on a socio-cognitive framework (Weir, 2005; O'Sullivan & Weir, 2011) combining social interpretations and personal experiences of English language tests (Yan & Horwitz, 2008) with cognitive dimensions of language use. Test-taker experiences were monitored and recorded before and after test-taking and twice during their first semester in university in 2014 (Figure 1). The focus of this paper was the institutional construct of a "task-in-process" (Seedhouse, 2005; Breen, 2009), that is, the interpretation of the test 'task' process rather than the test performance. Therefore, the evaluation (scoring) dimension of this framework was excluded.

Background

Pedagogy at the university English centre in this study focuses on helping students develop task-based academic English for university (Robinson, 2011) and the development of graduate attributes (Barrie, 2004). It is also a testing centre with responsibility to provide advice to students on the most appropriate English language test for their academic intention, learner type and career development. The four standardised English language proficiency tests currently accepted by most universities in Australia for international student entry purposes

were the subject of this study. They comprise The International English Language Testing System (IELTS), the Pearson Test of English-Academic (PTE-A), Cambridge English Advanced Exam (CAE), and Test of English as a Foreign Language internet-Based Test (TOEFL iBT). The first and third of these tests are paper-based, while the other two are computer-based. Three tests used in this research were 'live' official tests, and one, PTE-A, was an official practice test.

Participants

With the cooperation of teachers at the university English centre, the intent, duration and description of the research was first announced to students in class. Over 200 signed Expressions of Interest (EOIs) were collected from potential student participants at the university English centre with due confidentiality and anonymity precautions. The target cohort comprised full-time post-graduate and undergraduate university pathway students starting at the language centre in July 2013 and entering 13 faculties at the university in March 2014. They were representative of university pathway students in that they intended to enter university courses and had an English language achievement ranging from not less than IELTS 5.5, TOEFL iBT 65, PTE-Academic 68, CAE from Cambridge ESOL test scores 47. The purpose of the 25-week English centre university pathway program is to equip ESL students with appropriate English, study skills and cultural awareness required for tertiary study.

A meaningful sample size for this qualitative investigation was assessed to be over 10 participants due to triangulation of data sources, theory and participant review (Patton, 2002). As a result, 25 participants were considered to be a suitable size. In order to randomise the selection process, all names were entered in no particular order onto a spreadsheet and subject to the command {=RAND(first entry:last entry)}. 25 names were selected from the top, middle, and bottom of the randomised list. This reduced the pool to 75 potential participants. Those 75 names were transferred to another Excel spreadsheet and the process was repeated using 20 names from each part of the list. This pool of 60 was transferred to a new worksheet and again randomised. From this third list, the top 25 were selected as participants, and invited to a further briefing. The next 15 on the list were advised that they were 'standby' participants should

one of the original 25 decide to discontinue participation. The 15 'standby' volunteers were also invited to the next participant briefing.

The selected participants (*n* = 25) comprised 10 male and 15 female students aged 20 to 29 years, with 11 from China, 11 from Brazil, 2 two from Saudi Arabia and 1 one from Taiwan. In relation to previous test-taking experience, which was not a selection criterion, all of the non-Brazilian students had previously taken the IELTS at least once, and the student from Taiwan had taken the TOEFL paper-based test. Two language proficiency tests, the Cambridge CAE and the Pearson PTE-A, were unfamiliar to all participants.

Data Collection and Analysis

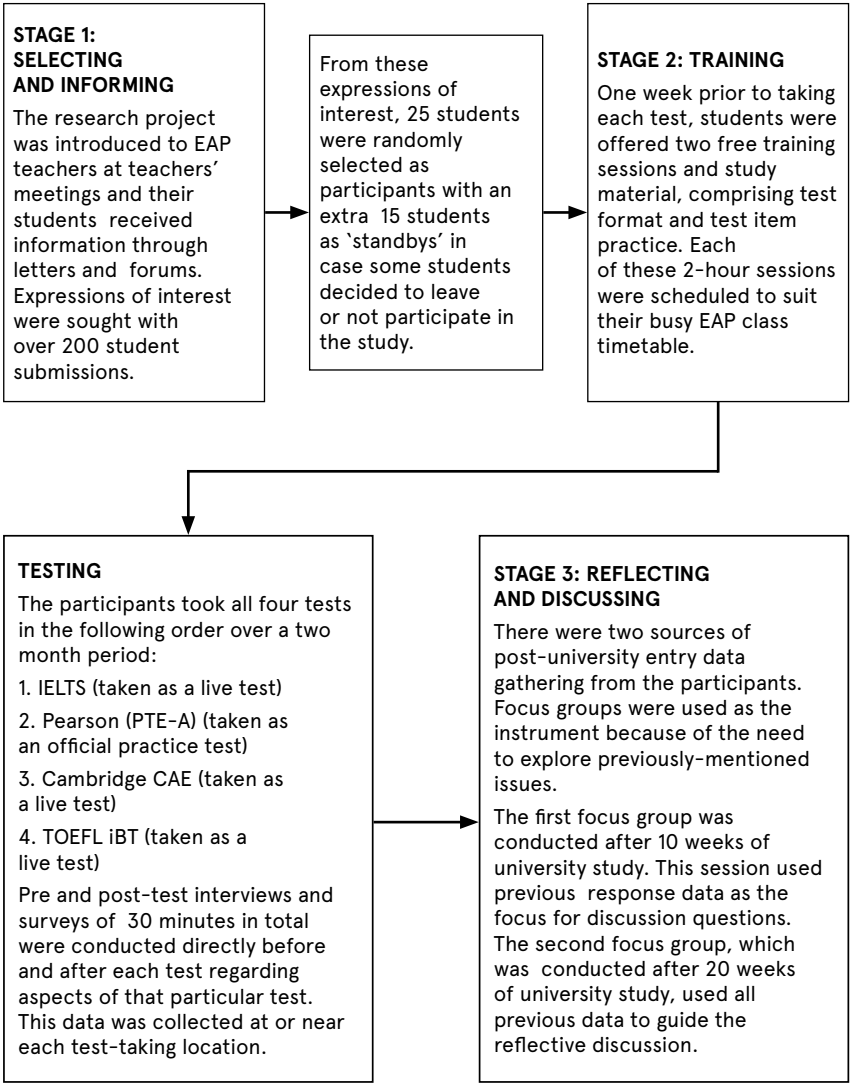
Through several qualitative sources of data-gathering, comprising open-ended questionnaires and two focus groups after entering university, student 'voices' were expressed before, during, and after each test between January and March 2014 (Figure 1). Questions at each stage of the data collection process were the same for all four tests. Included as part of the questionnaire, several Likert scale responses from strongly disagree [1] to strongly agree [5] provided a suitable quantitative measure of test-taking attitudes. One week prior to sitting each test, participants were offered two preparation classes of two hours each in that particular test for familiarisation purposes. Test familiarisation content for all four tests included test format and overview, test item examples and test item practice, with particular emphasis on less common items. PowerPoint slides and test practice materials were provided for students to review before taking each test. After entering university, participants were invited to discuss their test-taking experiences of each test again with the researchers. Focus groups were conducted after 10 weeks of university study and again after 20 weeks. At both these points in time, students met the researchers in several small groups to overcome conflicting study schedules. The questions in these post-university focus groups were more reflective in that they asked students to review their previous data as a source of "analytical triangulation" (Patton, 2002, p. 561).

Test-takers' perceptions were analysed using an inductive thematic coding process, followed by a deductive grouping of themes informed by previous studies. Firstly, student's responses were read

thoroughly to identify "recurring regularities" in the data (Patton, 2002, p. 465). Open codes were established to create tentative themes after reading all the data several more times. Labels were then attached to main ideas being expressed in the data. Possible categories were established by identifying recurring themes. This last process was repeated using keywords in context from the data to form new groupings. Miscellaneous data was revisited to determine if it matched any new categories. After colour-coding and data sorting those categories on the computer, responses were cut up and physically grouped to confirm themes and sub-theme clusters.

These inductively generated clustered themes and sub-themes emerging from the data were compared with previous test-takers' perspective studies (Cheng & DeLuca, 2011; Puspawati, 2012; DeLuca et al., 2013; Suryaningsih, 2014) in relation to nomenclature, relevance and consistency. Themes such as 'Scoring effects' (Cheng & DeLuca, 2011) and 'Perception on the score' (Puspawati, 2012) were considered irrelevant, and thus disregarded. Other themes in these studies such as 'Effects of the tests' (Suryaningsih, 2014), 'Psychological Factors' (Cheng & DeLuca, 2011), and 'Experience in taking the test' (Puspawati, 2012) seemed to overlap, so a more encompassing theme 'Affective factors' emerged. Likewise, 'Test content issues' emerged from similar themes in the four studies. Although two major themes labelled 'Testing environment' and 'Test design' (DeLuca *et al*., 2013, p. 670) developed from this process, sub-theme allocations from the four studies overlap seven different themes. Consequently, logical sub-theme clustering in our data determined theme allocation. The final theme, 'Idea development', was inductively generated from recurring comments and keywords such as 'brainstorming', 'synthesising', 'critical thinking' and 'thinking quickly.'

FIGURE 1
Stages of the test project



RESULTS

Five broad themes relating to the test-taking experience became apparent from the data: 1) testing environment, 2) test content issues, 3) affective factors, 4) test design, and 5) idea development. Results are organised according to these themes and key trends are illustrated using representative direct quotations from the raw data. An overall summary of the clustered theme results is provided in Figure 3 and macro-skill issues are shown in Figure 4 according to test-type.

Theme 1: Testing environment

The test-taking environment impacted on the test-takers in relation to instructions and noise (Figure 3). Regarding the sub-theme of 'instructions', students spoke about lexical complexity and not knowing when a section had finished in the PTE-A. This was especially true for reading task instructions, causing misunderstanding of some test items:

> *Because it is a proficiency test, [students at] any level should be able to understand the vocabulary in the instructions. (Focus Group 1)*

Noise-related disturbances were the only 'external factors' articulated by students in the feedback immediately after the PTE-A and TOEFL iBT tests and in subsequent focus groups. In the PTE-A exam room, a few nervous students apparently spoke loudly through their computer headphone speakers during the speaking component. This external noise was distracting for the other test-takers. The TOEFL iBT noise issue resulted from the extra tasks and length of some sections of the test. In other words, students were at different stages of the test, so some students were talking during the speaking section while others were concentrating on a different test section.

> *If you do a different part to another candidate such as an additional reading, you can hear people still doing the speaking test.*

Theme 2: Test content issues

Students experienced 'test-item difficulties' in all four tests, but the most commonly expressed problem related to extensive 'searching' for answers (Figure 3). Students perceived that the IELTS reading test comprised a high frequency of matching-type questions, ranging

from viewpoint-person matches to paragraph-heading matches. Participants also reported that a poor approach to 'Not Given' questions in the T/F/NG test items caused them to waste time searching for the answers. In the PTE-A, some task requirements were considered 'confusing' by test-takers. For example, multiple choice tasks requiring a single answer were sometimes followed by multiple choice tasks requiring several responses. Other aspects considered difficult in the PTE-A include converting notes to text in the listening test, typing, listening-memorisation, quickly describing graphs, summarising, and voice recording onto the computer. From several data sources, test-takers indicated the vocabulary used in this test is more academic and complex than in other exams.

> *I need to become familiar with academic words because the vocabulary in this test is a little bit more complex than the IELTS test.*

In the CAE test, vocabulary and language complexity was a recurring regularity in the data. Test-takers reported that different types and length of reading text as well as grammar-related questions added 'content' complexity to the exam. In the writing section, some students stated that because they were not confident or familiar with the report genre, it was problematic. While test-takers were overwhelmingly positive about speaking with a partner, common issues were topic development, language input, and strategy development. For the listening test, students considered various accents and conversations about everyday topics to be challenging at times. As one student commented,

> *It really tests the English knowledge of the student.*

In the TOEFL iBT, test-item content for listening and integrated test sections were commonly considered difficult by participants. In particular, they stated that transforming lecture notes into a summary is a useful but more difficult skill. Many students also reported the language and vocabulary of campus scenarios to be an issue in the listening and speaking test:

> *...knowing the vocabulary for campus conversations [is important]*

Theme 3: Affective factors

Anxiety levels before taking all four tests in response to the statement

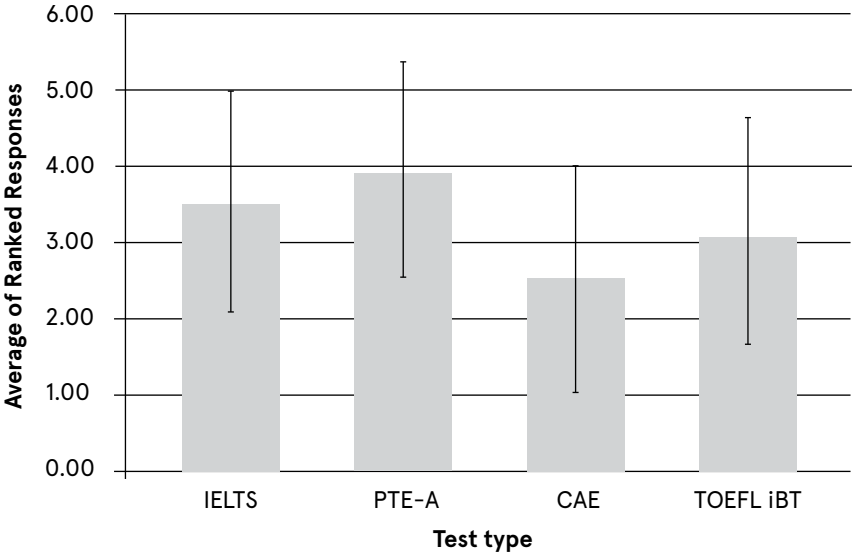
`I feel anxious before taking this test' were reported in a Likert-style item ranging from (1) I don't agree at all, to (5) I totally agree (Figure 2). The graph indicates that anxiety was a significant factor for test-takers in all tests, but the range of responses and sample size mean that it requires further investigation. Interestingly, one of the least familiar tests, the CAE test, recorded the lowest pre-test anxiety level (2.55, SD 1.293). Nevertheless, the high level of IELTS familiarity among participants from previous test-taking experiences generally resulted in positive affective factor responses. Students stated they were confident and all participants attended the preparation classes. However, overconfidence appeared to be a major issue as illustrated in the following comment.

I thought I did well, but the score is low (Focus Group 1)

Insufficient test preparation was articulated by participants as a concern for the other three tests, namely the PTE-A, CAE and TOEFL iBT (Figure 3). Attendance at these preparation sessions varied from fifty to eighty percent due to conflicting schedules. Participants unable to attend the sessions received Power Point slides and practice materials by email. The lack of familiarity before students took the PTE-A test may account for the pre-test anxiety level of 3.91 (SD 1.044) among test-takers (Figure 2). As the CAE speaking component follows a different format from the other tests, students were provided with a training video and review session. The following comments indicate that improved confidence was the perceived outcome for those who attended the preparation classes.

The classes were very good and helpful. I believe that without those classes I could not make the test as confident as I did. (Post-CAE test data)

FIGURE 2 PRE-TEST ANXIETY IN FOUR TESTS



Note: error bars represent standard deviation (SD)

Theme 4: Test design

The computer-based integrated test items in the PTE-A and TOEFL iBT tests received favourable participant feedback. However, the unfamiliar computer delivery mode and item design appeared to cause confusion in the PTE-A (Figure 3) as reflected in the following comment.

> *I felt nervous. Not because of the pressure from the teacher, but because the test is extremely different from how I expected. It is incomparable to any of the other English tests that I've done so far. It requires more than simple skills such as listening or reading. To do this test you must be really prepared, even to memorise short recordings.*

For most of the test-takers, the PTE-A was their first computer-based English proficiency test and they reported being confused about the mechanics and procedures of the test. Despite the high interest level in this test, students stated that the skills required were specific to PTE-A, not easily transferred from other tests, and

required much more than minimal practice for each new item type. Specialised training was also considered important in the TOEFL iBT test, especially for integrated tasks and online speaking responses. Even though this was their second computer-based speaking test (after PTE-A), they still experienced difficulty giving oral responses on the computer within the allocated time. One test-taker stated,

> *It was different from the traditional style so you need special training to do the test.*

The 'duration' of some tests and test sections received numerous comments in relation to fatigue. Test-takers remarked that for a computer-based test, the overall test duration and number of lectures in TOEFL iBT impacted on their concentration. One test-taker commented,

> *During the Reading Part, I felt a little bit tired because the exam was on the computer. So stay reading on the screen is harder to concentrate than reading a paper test.*

Several data sources mentioned the importance of listening in TOEFL iBT by the inclusion of this skill in both independent and integrated tasks. As one student reflected,

> *Lecture listening and note-taking is crucial for success.*

However, test-takers also experienced fatigue in the CAE paper-based test. The overall test time of 4 hours 40 minutes was considered too long and fatigue impacted on perceived confidence and performance levels. Furthermore, an emphasis on grammar and reading in the CAE was mentioned in different data sources. During the IELTS test, discomfort was an issue for some test-takers due to test continuation between sections. As the following comment illustrates,

> *Tests follow each other too quickly...cannot use [the] toilet.*

Theme 5: Idea Development

Test-takers identified aspects of all four tests that require idea development, including the sub-themes of 'quick thinking' (timing and response) and 'synthesising and critical thinking' (Figure 3).

Participants believed the IELTS writing section required the brainstorming of ideas followed by a quick essay plan. In other words, brainstorming was considered important in fulfilling test

requirements within the time constraints. Test-takers thought that this brainstorming problem also applied to Part 2 of the speaking test. Because this section of the speaking test requires candidates to think of ideas within 60 seconds for a 90-120 second talk, students stated the quick development of ideas is crucial. They noted that it was difficult to think of ideas quickly because the topics in Part 1 and Part 2 are unrelated. One student commented,

> *Speaking for 2 minutes about a topic requires creativity on an unfamiliar topic.*

Similarly, participants recognised that the PTE-A test required quick speaking responses for describing a visual such as a map or graph. After 15 seconds, candidates give a 45 second oral response. As one student stated,

> *I feel that this test has an interesting mode. This test requires quick reactions to what you've heard and what you've seen.*

They considered listening and reading items that tested global comprehension in PTE-A to be quite different from their previous test item experience, so strategies for this type of question were reported to be inadequate or underdeveloped.

> *To grasp the main idea rather than to catch every word was required.*

Higher-order thinking skills were reported as necessary in the CAE test to understand and answer reading questions. One student commented,

> *...it allows us to think and find the attitude of the writer and then to answer questions.*

Students remarked that integrated items in the TOEFL iBT test require higher-order language and thinking skills. This level of thinking also applied in the reading section, with students referring to critical thinking and referent tasks. Time constraints and quick responses were highlighted as difficult by participants for both speaking and writing. For instance, the essay writing requirement of up to 300 typed words within 30 minutes was reported as challenging for students with limited typing skills and for those used to tests with different time and length constraints.

Figure 3 shows the themes and sub-themes generated in this study. The most common problem type for each sub-theme is identified

along with cross-data occurrence in the four tests. Illustrations of each problem are provided through direct quotes from the raw data.

FIGURE 3.
Clustered theme summary chart

Themes and sub-themes	Problem type	Test type				Illustrations
		IELTS	PTE-A	CAE	TOEFL -IBT	
Theme 1 Testing environment						
instructions	Complex instructions		◆◆			PTE-A: 'PTE is very complicated and the instruction words are hard'
external factors (noise)	Disturbed by noise during the test		◆◆		◆◆	PTE-A: 'I couldn't concentrate because everyone in class was speaking at the same time' TOEFL iBT: 'If you do a different part to another candidate such as an additional reading, you can hear people still doing the speaking test.'
Theme 2 Test content issues						
test item difficulties	Some item types require a lot of 'searching'	◆	◆	◆	◆	PTE-A: 'I had a problem with this test that was memorizing long sentences and write them later.' IELTS: 'Searching for NG answers wasted a lot of time'

content difficulties	Vocabulary and texts are at a high level		◆◆	◆◆	◆	PTE-A: `...frustrating because it's too difficult if your skill level is low' CAE: `...very advanced with professional language'; `the text is more complex, like the texts in uni' TOEFL iBT: `Topics are organised well but it requires a higher level of vocabulary knowledge.'
Theme 3 Affective factors						
anxiety	Nervous before testing and during answer searches	◆◆	◆◆	◆◆	◆◆	IELTS: `A bit nervous before but OK during [the] test'; `Tried to keep calm but in [the] reading test I didn't find the answer and became nervous'
familiarity	Test items are unfamiliar		◆◆	◆	◆◆	`The accumulated experience and knowledge comes from the preparation classes and [taking] the four exams.'
preparation	Insufficient test preparation		◆	◆	◆	CAE: `I believe that without those classes I could not make the test as confident as I did.' PTE-A: `To do this test you must be really prepared, even to memorise short recording[s].'

confidence	Confidence is over / under-estimated	◆		◆		IELTS:'I thought I did well, but the score is low' 'Test scores do not represent English confidence'
Theme 4 Test design					
integrated items	unfamiliar mode and item design		◆◆		◆	PTE-A: 'I felt a little bit confused because this test differ[s] a lot from the others.' TOEFL iBT: 'Adapting to the new type of questions was difficult.'
test mode	-Unused to speaking onto a computer -Typing takes time		◆		◆	'It would be good to speak with a person instead of a computer' 'Some [candidates] have an advantage in typing skills' TOEFL iBT: 'I love the integrated questions but I'm not used to speaking with a computer'
duration (and fatigue)	Test length caused fatigue or discomfort	◆		◆◆	◆◆	TOEFL iBT: ' Too long for a computer-based test' ;'I got a headache from using headphones with glasses during the computer tests' CAE: 'Tests are too long, especially the CAE' IELTS: 'Tests follow each other too quickly... cannot use[the] toilet'

test structure	Some test sections given extra emphasis			◆◆	◆◆	TOEFL iBT: 'Listening was too long- lectures were difficult.' CAE: 'A good test for grammar and listening- especially different accents.... but it's hard!'
Theme 5 Idea development						
quick thinking (timing and response)	Ideas need developing quickly	◆◆	◆	◆	◆	IELTS: 'Speaking for 2 minutes about a topic requires creativity on an unfamiliar topic'; 'Time was limited, especially [in the] reading test' TOEFL iBT: 'Speaking is difficult because 30 to 40 seconds preparation is too short. It's not enough time to react.'
synthesising and critical thinking	Higher- level thinking can be difficult	◆	◆	◆◆	◆◆	CAE: ...'it allows us to think and find the attitude of the writer and then to answer questions' PTE-A: 'To grasp the main idea rather than to catch up with every word was required.'

Note: One diamond (◆) indicates data from a single source and two diamonds (◆◆) indicate data from more than one source

In Figure 4, the data has been extracted, reorganised and divided according to skill-type. The purpose of this figure is to indicate tasks that students reported as challenging in the four tests. This chart could help inform those who have a responsibility to prepare the test-takers. As one student requests,

According to our different learning skill, find the weak point and help us overcome it.

FIGURE 4.
Student perceptions of problematic aspects of four standardised tests

Test type	Receptive skills		Productive skills		Mixed skills
	Listening	Reading	Speaking	Writing	Integrated
IELTS	· Reading instruction in a limited time · Knowing everyday expressions · Answering matching- type questions · Spelling correctly	· Answering questions in a limited time · Dealing with NG questions · Matching-type questions · Scanning for details	· Thinking quickly of ideas for Part 2 long turn (Part 1 and Part 2 not thematically connected)	· Timing for Task 1 and Task 2 · Thinking quickly of ideas for Task 2 essay	· Not applicable
PTE-A	· Applying strategies for less familiar tasks e.g. notes to text · Understand-ing instruction vocabulary	· Understanding the task requirements e.g. different types of multiple choice · Knowing strategies for unfamiliar item types	· Describing a graph or visual in a limited time · Speaking onto a computer (ignoring other speakers)	· Timing skills	· Typing and memorisation skills · Typing quickly · Knowing strategies for integrated tasks e.g. summary
CAE	· Becoming familiar with a wide range of accents · Listening to everyday conversa-tions	· Reading widely · Reading instructions quickly · Understanding detailed grammar	· Knowing how to communi-cate with a partner	· Using the report genre · Timing for essay writing	· Not applicable

TOEFL iBT	· Having a range of vocabulary for campus topics · Using note-taking skills	· Reading academic-style text · Thinking critically · Completing referent tasks e.g. 'it' ; 'they'	· Responding to questions quickly · Using vocabulary for campus topics	· Expressing clear opinion on an issue · Timing the 300 word essay	· Transforming notes to a summary · Typing skills · Thinking critically

DISCUSSION

The aim of this study was to investigate how student perceptions of four university gateway tests (IELTS, PTE-A, CAE and TOEFL iBT) could better inform educators at university English centres. Although we acknowledge that experiences from 25 test-takers cannot lead to generalisations, they do offer insights into areas for further investigation. The discussion will explore how these results could apply to the development of appropriate test preparation courses or even be integrated into other courses by discussing paper-based tests followed by computer-based tests related to each research question.

Regarding the first research question (How do students perceive their experiences of taking four university gateway English proficiency tests?), it appears that affective factors such as test anxiety (Winke & Lim, 2014) have an impact on the student's test taking experience. Reported anxiety during the IELTS exam mainly consisted of three aspects: time constraints in writing, difficulty finding answers in the reading section and thinking of ideas in the speaking section (Figure 4). However, the students were positive about their test-taking experiences regardless of any previous test score. If preparation courses could address the three aforementioned areas that cause student anxiety in the IELTS, they might improve test outcomes. Furthermore, except for the reported 'cannot use toilet' discomfort issue in the current study, students confirmed other results by Suraningsih (2014) that test-fatigue and noise were not issues. Student anxiety for the CAE, the other paper-based test, may relate to the inherent difficulty of the test rather than any test design issues. Even a short preparation course was perceived to be helpful in relieving anxiety (Liu, 2014), but all students stated that a longer time is needed to adequately prepare for this advanced level test. Overwhelming positive experiences from all participants were

perceived for the CAE interactive speaking section, and this positive feeling was unrelated to their test scores. Other positively perceived experiences were the writing task options and highlighting function.

The reported pre-test anxiety level for PTE-A (Figure 2) was higher than for other tests, possibly linked to the large number of unfamiliar test items and the unfamiliar computer test mode. Breen (2009) states that learners only relate to tasks through what they recognise as familiar. This importance of task familiarity is supported in the current study where students perceived that the specific skills required for this test could not be easily transferred from other test-taking experiences. In this study, 'external factor' results in both computer-based tests (Figure 3) confirmed previous research results (Puspawati, 2012; Suraningsih, 2014; DeLuca *et al*., 2013) reporting that noise from other test-takers disturbed candidates. Perceptions of headphone discomfort and test duration fatigue in this study also confirmed those reported by De Luca et al.'s (2013) TOEFL iBT study. However, the reported anxiety level of participants before the TOEFL iBT (Figure 2) was lower than expected for a test with both unfamiliar test-mode and test items, indicating further research is necessary. Topic familiarity or topic-specific vocabulary was perceived to be an issue to some degree in all tests, which was also reported in other research (He & Shi, 2008; Puspawati, 2012; DeLuca *et al*., 2013, Suraningsih, 2014).

Regarding the second research question (What test-item types and skill types do test-takers perceive to be more challenging across these four high-stakes tests, thus requiring more preparation?), a number of areas of consensus were found among participants. By listening to students 'voice' in this area, teachers can ensure that negative factors are minimised, making test scores more valid (Xie & Andrews, 2012). Although the two paper-based tests, IELTS and CAE, had a more familiar delivery mode, some aspects were perceived as challenging. Test-takers considered some types of questions in the IELTS problematic due to time constraints. 'Not Given' responses and many matching-type questions, for example, apparently caused test-takers to 'waste' a lot of valuable test time, with serious consequences from the student's perspective (Figure 3). In the IELTS speaking component, students thought it was difficult to transition from Part 1 to Part 2 (the long turn)

because of a lack of context, making idea development challenging. This topic development issue was also found in a similar study by Suraningsih (2014). The CAE is a different type of test from the other three in that it is an advanced test and not designed to be taken at all English proficiency levels. Indeed, due to its intended purpose, it would be a problem if the students considered it to be easy. Nevertheless, extensive reading skills and familiarity with a range of accents were seen as specific skills required for this test. Timing, grammar knowledge, report writing genre and interacting with a partner were perceived by the students as areas that were difficult and would require further attention. One limitation of the current study is that participants took the 2014 CAE test which has since been amended. While some results may be different for the 2015 version, most should still be relevant. The revised test coincidentally implemented student's ideas such as merging the 'Use of English' grammar section to create an overall shorter test.

Aspects of standardised tests perceived as challenging include unfamiliar tasks often linked to unfamiliar delivery modes. Regarding insufficient familiarity of task and delivery mode, students perceived the most difficulty in Pearson's PTE-A and TOEFL iBT. A number of unfamiliar tasks in PTE-A such as various types of multiple choice, repeating a sentence orally, describing a visual graphic orally or summarising in one sentence were further complicated by less common task instruction lexicon according to student feedback (Figure 3). In the TOEFL iBT, understanding lexicon based on campus scenarios was a perceived difficulty along with listening-note taking skills, but Webb and Paribakht (2015) warn against corpus-driven lexical comprehension assumptions based on a particular discourse type. In particular, students thought it was challenging to transform their bullet-pointed notes into a cohesive summary and that listening skills comprised a disproportionately large component of the test. In addition, both of these tests are computer-based and the 'mechanics' or procedures of these computer-based tests, including typing and speaking requirements, could cause more challenges for some students than in other tests. Surprisingly, even though it was pointed out in group focus discussions that students have to use computers and other technological devices on a daily basis, most maintained their

position on this issue. In fact, only the two Saudi students spoke in favour of typed responses and speaking onto a computer. It appears that computer familiarity levels from their home countries may have long-term impacts on test-mode attitudes, perhaps contributing to the negative speaking test perceptions noted by Stricker and Attali (2010) and the typing effect reported by Barkaoui (2014). Indeed, test content rather than test-mode or affective factors seemed to change positively over time. This observation requires further research to determine whether test-mode and other attitudes are mutable or not. While integrated test items were often considered by students to be more linguistically challenging (Deluca *et al*., 2013), it was felt the academic rewards of such tasks flow on to their university study (as conveyed in both focus groups). This finding supports the inclusion of integrated skills in standardised tests because when language skills are perceived as necessary by test-takers to correctly answer items, Xie and Andrews (2012) assert it is an endorsement of the test design.

Implications for teaching

Test preparation literature is often polemical (Reich & Bally, 2010). Whereas Green (2013) and Hughes (1989) advocate language skills development covering a wide range of skills, Liu (2014) contends that more intensive preparation courses may yield better short-term results. Moore and Morton (2005) recommended test preparation and EAP programs as separate courses due to differences between IELTS and university writing genres. However, the current study indicates that the decision to integrate or not may be more complex. For instance, the university skills required in EAP include language and study skills development, which are also aspects of tests. Furthermore, a number of skills identified as problematic such as transferring notes to text, report writing, campus scenario vocabulary, describing graphs and thinking critically are also EAP constructs. Finally, the degree of test familiarity and type of problematic items or skills could determine the length and type of separate short preparation required.

For IELTS, a General English course supplemented by a short preparation course may be sufficient preparation for students familiar with the test. As Winke and Lim (2014) concluded, even short preparation courses focusing on test format can reduce anxiety. They

also noticed anxiety related to time spent reading instructions and searching for key words, a similar processing problem identified in this study to 'Not Given' searches (Figure 4). Students suggested that Part 2 speaking anxiety could be alleviated by recording on mobile phones, building interactivity, self-efficacy, critical listening, and improved language skills. CAE test preparation could be included in high-level General English courses alongside separate intensive courses (Liu, 2014). To address critical thinking and high-level reading skills, Fahim and Tabataba'ian (2012, p. 132) recommended 'concept mapping' for activating a wider range of reading strategies through visual representation. For the long essay, lesson time could be allocated for students to develop clear opinions on an issue within a limited time on the computer. As Pearson's PTE-A included unfamiliar test items, a longer separate course may be necessary. However, perceived problematic items such as summaries, graph description and typed responses, could be taught in an EAP course. Memory was mentioned by participants in the current study as a factor impacting on their listening in integrated items. A study by Brunfaut and Revesz (2015) on L2 listening and anxiety also found lower listening performance corresponded significantly with lower memory capacity and anxiety. He and Shi (2008) further noted that memory was considered important by Chinese test-takers, indicating a need for further investigation into memory and testing. The TOEFL iBT appears aligned to EAP construct (DeLuca *et al*., 2013) which could address note-taking, summary writing, academic reading, typing, critical thinking, and campus vocabulary. Moreover, Matoush and Fu (2012) claim college life 'enrichment courses' could build campus scenario vocabulary. Participants highlighted the need for well-developed listening skills for integrated and independent tasks. This skill could be addressed using free computer software to slow audio tempo without impacting on pitch (East & King, 2012).

CONCLUSION

This paper has addressed a common dilemma faced by university English centres in relation to preparing and advising students taking a university gateway language test. A discussion of students' perceptions of the test-taking experience was linked with current theory and important implications for TESOL practice. This could

provide a clearer direction for TESOL practitioners at university English centres who are given the imperative to prepare students for such tests. The exploratory nature of the findings presented in this paper illustrates the need for further investigation including the interrelationships between preparation, test type and performance.

Aspects in four tests perceived by students as being problematic have been identified. Generally, students reported that receptive and mixed skills were more problematic than productive skills. In particular, integrated test items were considered to be more challenging but also more academically rewarding. 'Idea development' was a theme reported as causing difficulty, indicating a possible role for 'test-management skills' in preparation courses. As expected, test format and test item familiarity were considered crucial in reducing anxiety and improving test score validity. If courses could help familiarise students with the format, item types, and skills reported as 'difficult' in this paper, it would be likely to improve score reliability and student outcomes.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support and financial assistance of Educational Testing Service (ETS), Cambridge, Pearson, and the Centre for English Teaching (CET) at the University of Sydney. We also thank the participating students for their commitment to the study, and the anonymous reviewers for their valuable feedback on an earlier draft.

THE AUTHORS

John Gardiner is a teacher at the Centre for English Teaching, University of Sydney. He received a Master's degree in TESOL from Bond University and a Diploma of Teaching (Primary) from the University of New South Wales. He has spent the last 30 years in a diverse range of educational contexts both in Australia and overseas. His research interests include language testing, action research, curriculum design and course development.

Qualified EdD and MEd in international education, and BAdultEd in vocational education, Stephen Howlett has spent the past 15 years as a lecturer and administrator in various aspects of international education, management and leadership in the Sydney area.

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Ata, A. W. (2015). Knowledge, education, and attitudes of international students to IELTS: A case of Australia. *Journal of International Students*, 5(4), 488-500.

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241-259.

Barrie, S. C. (2004). A research-based approach to generic graduate attributes policy. *Higher Education Research & Development, 23*(3), 261-275.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377.

Breen, M.P. (2009). Learner contributions to task design. In K. Van den Brandon, M. Bygate & J.M. Norris (Eds.), *Task-based language teaching: A reader* (pp. 333-356). Philadelphia, PA: John Benjamins Publishing Company.

Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly, 49*(1), 141-168.

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104-122.

Cheng, L. Y., Sun, Y. Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436-470.

Cohen, A. D. (2013). Using test-wiseness strategy research in task development. In A.J. Kunnan (Ed.), *The companion to language assessment.* (pp. 893-905). Hoboken, NJ: Wiley/Blackwell.

DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: an exploratory study on the TOEFL iBT. *System, 41*(3), 663-676.

Dunworth, K. (2010). Clothing the emperor. *Australian Universities'*

Review, 55(2), 5-10.

East, M., & King, C. (2012). L2 learners' engagement with high stakes listening tests: Does technology have a beneficial role to play? *CALICO Journal, 29*(2), 208-223.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *International English Language Testing System (IELTS) Research Reports, 4*(6), 207-254.

Fahim, M., & Tabataba'ian, M. S. (2012). Concept maps, cloze tests, and multiple-choice tests: A think-aloud approach to the comparison of the strategies utilised in different test formats. *Journal of American Science, 8*(8), 131-138.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education, 14*(1), 75-97.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies, 13*(2), 39-51.

He, L., & Shi, L. (2008). ESL students' perceptions and experiences of standardised English writing tests. *Assessing Writing, 13*(2), 130-149.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Liu, O. L. (2014). Investigating the relationship between test preparation and TOEFL iBT performance. *ETS Research Report Series*, (2), 1-13.

Matoush, M. M., & Fu, D. (2012). Tests of English language as significant thresholds for college-bound Chinese and the washback of test-preparation. *Changing English, 19*(1), 111-121.

McNeil, L. (2011). Investigating the contributions of background knowledge and reading comprehension strategies to L2 reading comprehension: An exploratory study. *Reading and Writing, 24*(8), 883-902.

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*(1), 43-66.

Moore, W. P. (1994). Appropriate test preparation: Can we reach a consensus? *Educational Assessment, 2*(1), 51-68.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146-160.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363-380.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13-32). Basingstoke, UK: Palgrave Macmillan.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12-15.

Puspawati, I. (2012). *EFL/ESL (English as a Foreign/Second Language) Students' Perceptions toward the TOEFL (Test of English as a Foreign Language) Test* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. UMI 1515041)

Reich, G. A., & Bally, D. (2010). Get smart: Facing high-stakes testing together. *The Social Studies, 101*(4), 179-184.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*(s1), 1-36.

Sadeghi, S. (2014). High-stake test preparation courses: Washback in accountability contexts. *Journal of Education & Human Development, 3*(1), 17-26.

Seedhouse, P. (2005). "Task" as research construct. *Language Learning, 55*(3), 533-570.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* London: Pearson.

Song, X., & Cheng, L. (2006). Language learner strategy use and test performance of Chinese learners of English. *Language Assessment Quarterly: An International Journal, 3*(3), 243-266.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence:

A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747-758.

Stricker, L. J., & Attali, Y. (2010). Test takers' attitudes about the TOEFL IBT. *ETS Research Report Series*, (1), i-16.

Suryaningsih, H. (2014). *Students' perceptions of international English language testing system (IELTS) and test of English as a foreign language (TOEFL) tests* (Doctoral dissertation). Indiana University of Pennsylvania, Indiana, PA. Retrieved from ProQuest Dissertations and Theses. (Accession Order No. UMI 1555587)

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardised English proficiency test? *English for Specific Purposes, 38*(1), 34-43.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Winke, P., & Lim, H. (2014). Effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation, *IELTS Research Reports Online Series, 3*(1), 1-30. Retrieved from http://www.ielts.org/researchers

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49-70.

Yan, J. X., & Horwitz, E. K. (2008). Learners' perceptions of how anxiety interacts with personal and instructional factors to influence their achievement in English: A qualitative analysis of EFL learners in China. *Language Learning, 58*(1), 151-183.

Yu, G. (2012). *Preparing for TOEFL iBT speaking tasks: test-takers' experiences and expectations*. Paper presented at the 34th Language Testing Research Colloquium, Princeton, New Jersey.